P /DE 99/01323

BUNDESREPUBLIK DEUTSCHLAND

200143

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

DE 99/1323

EU

Bescheinigung

Die Siemens Aktiengesellschaft in München/Deutschland hat eine Patentanmeldung unter der Bezeichnung

"Anordnung und Verfahren zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache durch einen Rechner"

am 11. Mai 1998 beim Deutschen Patent- und Markenamt eingereicht.

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

Die Anmeldung hat im Deutschen Patent- und Markenamt vorläufig das Symbol G 10 L 5/06 der Internationalen Patentklassifikation erhalten.

München, den 15. Juni 1999

Deutsches Patent- und Markenamt

Der Präsident

Im Auftrag

Agurks

Aktenzeichen: 198 21 057.4

A 9161
06.90
11/98

198 21 057.4 vom 11.5.98

Beschreibung

Anordnung und Verfahren zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache durch einen Rechner

Die Erfindung betrifft eine Anordnung und ein Verfahren zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache durch einen Rechner.

Ein Verfahren und eine Anordnung zur Erkennung gesprochener Sprache sind aus [1] bekannt. Bei der Erkennung gesprochener Sprache werden, insbesondere bis zum Erhalt einer erkannten Wortfolge aus einem digitalisierten Sprachsignal, eine Signalanalyse und eine globale Suche, die auf ein akustisches Modell und ein linguistisches Modell der zu erkennenden Sprache zurückgreift, durchgeführt. Ein akustisches Modell basiert auf einem Phoneminventar, das anhand von Hidden-Markov-Modellen (HMMs) realisiert ist. Während der globalen Suche werden für Merkmalsvektoren, die aus der Signalanalyse hervorgegangen sind, mit Hilfe des akustischen Modells eine passende Wortfolge ermittelt und diese als erkannte Wortfolge ausgegeben. Die zu erkennenden Wörter sind in einem Aussprachelexikon zusammen mit einer phonetischen Umschrift abgespeichert. Der Zusammenhang ist ausführlich in [1] dargestellt.

Zur Erläuterung der nachfolgenden Ausführungen wird an dieser Stelle kurz auf die verwendeten Begriffe eingegangen.

Die Signalanalyse als Phase der computerbasierten Spracherkennung umfaßt insbesondere eine Fouriertransformation des digitalisierten Sprachsignals und eine sich daran anschließende Merkmalsextraktion. Aus [1] geht hervor, daß die Signalanalyse alle zehn Millisekunden erfolgt. Aus sich überlappenden Zeitabschnitten mit einer Dauer von z.B. jeweils 25 Millisekunden werden anhand der Signalanalyse ungefähr 30 Merkmale ermittelt und zu einem

Merkmalsvektor zusammengefaßt. Die Komponenten des Merkmalsvektors beschreiben die spektrale Energieverteilung des zugehörigen Signalausschnitts. Um diese Energieverteilung zu erhalten, wird auf jedem Signalabschnitt (25ms-Zeitabschnitt) eine Fouriertransformation durchgeführt. Aus der Darstellung des Signals im Frequenzbereich resultieren die Komponenten des Merkmalsvektors. Nach der Signalanalyse liegt das digitalisierte Sprachsignal in Form von Merkmalsvektoren vor.

Diese Merkmalsvektoren werden der globalen Suche, einer weiteren Phase der Spracherkennung, zugeführt. Wie bereits erwähnt, bedient sich die globale Suche des akustischen Modells und ggf. des linguistischen Modells, um die Folge von Merkmalsvektoren auf Einzelteile der als Modell vorliegenden Sprache (Vokabular) abzubilden. Eine Sprache setzt sich aus einer vorgegebenen Anzahl vom Lauten, sog. Phonemen, zusammen, deren Gesamtheit als Phoneminventar bezeichnet wird. Das Vokabular wird durch Phonemfolgen modelliert und in einem Aussprachelexikon abgespeichert. Jedes Phonem wird durch mindestens ein HMM modelliert. Mehrere HMMs ergeben einen stochastischen Automaten, der Zustände und Zustandsübergänge (Transitionen) umfaßt. Mit HMMs läßt sich der zeitliche Ablauf des Auftretens bestimmter Merkmalsvektoren (selbst innerhalb eines Phonems) modellieren. Ein entsprechendes Phonem-Modell umfaßt dabei eine vorgegebene Anzahl von Zuständen, die linear hintereinander angeordnet sind. Ein Zustand eines HMMs stellt einen Teil eines Phonems (bspw. mit einer Dauer von 10ms) dar. Jeder Zustand ist verknüpft mit einer Emissionswahrscheinlichkeit, die insbesondere nach Gauß verteilt ist, für die Merkmalsvektoren und mit Transitionswahrscheinlichkeiten für die möglichen Übergänge. Mit der Emissionsverteilung wird einem Merkmalsvektor eine Wahrscheinlichkeit zugeordnet, mit der dieser Merkmalsvektor in einem zugehörigen Zustand beobachtet wird. Die möglichen Übergänge sind ein direkter Übergang von einem Zustand in

einen nächsten Zustand, ein Wiederholen des Zustands und ein Überspringen des Zustands.

Eine Aneinanderreihung von HMM-Zustände mit den zugehörigen Übergängen über die Zeit wird als Trellis bezeichnet. Um die akustische Wahrscheinlichkeit eines Wortes zu bestimmen, verwendet man insbesondere das Prinzip der dynamischen Programmierung: Es wird der Pfad durch die Trellis gesucht, der den geringsten Fehler aufweist bzw. der durch die größte Wahrscheinlichkeit für ein zu erkennendes Wort bestimmt ist.

Das Ergebnis der globalen Suche ist die Ausgabe bzw. Bereitstellung einer erkannten Wortfolge, die sich unter Berücksichtigung des akustischen Modells (Phoneminventar) für jedes einzelne Wort und des Sprachmodells für die Abfolge von Wörtern ergibt.

Aus [2] ist ein Verfahren zur Sprecheradaption, basierend auf einer MAP-Schätzung (MAP = maximum a posteriori) von HMM-Parametern bekannt.
So ist es laut [2] anerkannt, daß ein sprecherabhängiges System zur Spracherkennung normalerweise bessere Ergebnisse als ein sprecherunabhängiges System liefert, sofern ausreichend Trainingsdaten verfügbar sind, die eine Modellierung des sprecherabhängigen Systems ermöglichen. Sobald jedoch die Menge der sprecherspezifischen Trainingsdaten beschränkt ist, erreicht das sprecherunabhängige System die besseren Resultate. Eine Möglichkeit zur Leistungssteigerung beider Systeme, also sowohl des sprecherabhängigen als auch des sprecherunabhängigen Systems zur Spracherkennung, besteht darin, die vorab gespeicherten Datensätze mehrerer Sprecher, derart zu benutzen, daß auch eine kleine Menge Trainingsdaten ausreicht, um einen neuen Sprecher in ausreichender Qualität zu modellieren. Solch ein Trainingsverfahren wird Sprecheradaption genannt. In [2] wird insbesondere die

Sprecheradaption durch eine MAP-Schätzung der Hidden-Markov-Modell-Parameter durchgeführt.

Generell verschlechtern sich Ergebnisse eines Verfahrens zur Erkennung gesprochener Sprache, sobald charakteristische Merkmale der gesprochenen Sprache von charakteristischen Merkmalen der Trainingsdaten abweichen. Beispiele für charakteristische Merkmale sind Sprechereigenschaften oder akustische Kontexte, die sich in Form von Verschleifungen auf die Artikulation der Phoneme auswirken.

Der in [2] verfolgte Ansatz zur Sprecheradaption besteht darin, Parameterwerte der Hidden-Markov-Modelle "nachzuschätzen", wobei diese nach Verarbeitung "offline", d.h. nicht zur Laufzeit des Verfahrens zur Spracherkennung, durchgeführt wird.

Die **Aufgabe** der Erfindung besteht darin, eine Anordnung und ein Verfahren zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache anzugeben, wobei insbesondere eine Anpassung des akustischen Modells zur Laufzeit (also "Online") vollzogen wird.

Diese Aufgabe wird gemäß den Merkmalen der unabhängigen Patentansprüche gelöst.

Zur Lösung der Aufgabe wird ein Verfahren zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache durch einen Rechner angegeben, in dem aus der gesprochenen Sprache ein Sprachsignal bestimmt wird. Das Sprachsignal wird einer Signalanalyse unterworfen, woraus Merkmalsvektoren zur Beschreibung des digitalisierten Sprachsignals hervorgehen. Eine globale Suche wird zur Abbildung der Merkmalsvektoren auf eine in modellierter Form vorliegende Sprache durchgeführt, wobei jedes Phonem der Sprache durch ein modifiziertes Hidden-Markov-Modell und jeder Zustand des modifizierten Hidden-Markov-Modells durch eine

Wahrscheinlichkeitsdichtefunktion beschrieben wird. Es erfolgt eine Anpassung der Wahrscheinlichkeitsdichtefunktion derart, daß sie in eine erste Wahrscheinlichkeitsdichtefunktion und in eine zweite Wahrscheinlichkeitsdichtefunktion aufgespalten wird. Schließlich wird von der globalen Suche eine erkannte Wortfolge bereitgestellt.

Hierbei sei angemerkt, daß die Wahrscheinlichkeitsdichtefunktion, die in eine erste und in eine zweite Wahrscheinlichkeitsdichtefunktion aufgespalten wird, eine Emissionsverteilung für einen vorgegebenen Zustand des modifizierten Hidden-Markov-Modells darstellen kann, wobei diese Emissionsverteilung auch eine Überlagerung mehrerer Wahrscheinlichkeitsdichtefunktionen, z.B. Gauß-Kurven (Gauß'sche Wahrscheinlichkeitsdichteverteilungen), enthalten kann.

Eine erkannte Wortfolge kann dabei auch einzelne Lauten bzw. nur ein einzelnes Wort umfassen.

Sollte im Rahmen der globalen Suche eine Erkennung mit einem hohen Wert für den Abstand zwischen gesprochener Sprache und von der globalen Suche ermittelten dazugehöriger Wortfolge behaftet sein, so kann die Zuordnung eines Nullwortes erfolgen, welches Nullwort anzeigt, das die gesprochene Sprache nicht mit ausreichender Güte erkannt wird.

Es ist ein Vorteil der Erfindung, durch die Aufspaltung der Wahrscheinlichkeitsdichtefunktion in einem durch die Merkmalsvektoren aufgespannten Merkmalsraum neue Bereiche zu schaffen, die signifikante Information in Bezug auf die zu erkennenden digitalisierten Sprachdaten aufweisen und damit eine verbesserte Erkennung zu gewährleisten.

Eine Ausgestaltung besteht darin, daß die Wahrscheinlichkeitsdichtefunktion in die erste und in die zweite Wahrscheinlichkeitsdichtefunktion aufgespalten wird,

falls der Abfall eines Entropiewertes unterhalb einer vorgegebenen Schranke liegt.

Die Aufspaltung der Wahrscheinlichkeitsdichtefunktion in Abhängigkeit von einem Entropiewert erweist sich in der Praxis als äußerst vorteilhaft.

Die Entropie ist allgemein ein Maß für eine Unsicherheit bei einer Vorhersage eines statistischen Ereignisses. Die Entropie ist insbesondere mathematisch bestimmbar für Gauß-Verteilungen, wobei eine direkte logarithmische Abhängigkeit zwischen der Streuung σ und der Entropie besteht.

Eine andere Ausgestaltung der Erfindung besteht darin, daß die Wahrscheinlichkeitsdichtefunktionen, insbesondere die erste und die zweite Wahrscheinlichkeitsdichtefunktion jeweils mindestens eine Gauß-Verteilung umfassen.

Die Wahrscheinlichkeitsdichtefunktion des Zustandes wird durch eine Summe mehrerer Gaußverteilungen angenähert. Die einzelnen Gaußverteilungen werden Moden genannt. Bei dem vorgestellten Verfahren werden die Moden insbesondere isoliert voneinander betrachtet. Bei jedem einzelnen Aufspaltvorgang wird eine Mode in zwei Moden aufgeteilt. Wenn die Wahrscheinlichkeitsdichtefunktion aus M Moden gebildet wurde, so wird sie nach dem Aufspaltvorgang aus M+1 Moden gebildet. Wird eine Mode beispielsweise als eine Gaußverteilung angenommen, so kann eine Entropie berechnet werden, wie im Ausführungsbeispiel gezeigt wird.

Eine Online-Adaption ist deshalb vorteilhaft, weil das Verfahren nach wie vor Sprache erkennt, ohne in einer gesonderten Trainingsphase auf die Veränderung des Wortschatzes eingestellt werden zu müssen. Es erfolgt eine Selbstadaption, die insbesondere notwendig wird durch eine veränderte Koartikulation der Sprecher aufgrund eines Hinzufügens eines neuen Wortes.

Die Online-Adaption erfordert demnach keine gesonderte Berechnung der Wahrscheinlichkeitsdichtefunktionen, die wiederum für eine Nicht-Verfügbarkeit des Systems zur Spracherkennung verantwortlich wäre.

Eine Weiterbildung der Erfindung besteht darin, daß für die erste Wahrscheinlichkeitsdichtefunktion und für die zweite Wahrscheinlichkeitsdichtefunktion gleiche Standardabweichungen bestimmt werden. Ein erster Mittelwert der ersten Wahrscheinlichkeitsdichtefunktion und ein zweiter Mittelwert der zweiten Wahrscheinlichkeitsdichtefunktion werden derart bestimmt, daß der erste Mittelwert von dem zweiten Mittelwert verschieden ist.

Dies ist ein Beispiel für die Gewichtung der aus der Wahrscheinlichkeitsdichtefunktion aufgespaltenen ersten und zweiten Wahrscheinlichkeitsdichtefunktion. Es sind auch beliebig andere Gewichtungen vorstellbar, die auf den jeweiligen Anwendungsfall anzupassen sind.

Schließlich ist es eine Weiterbildung, daß das Verfahren mehrfach hintereinander durchgeführt wird und somit eine wiederholte Aufspaltung der Wahrscheinlichkeitsdichtefunktion erfolgt.

Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Eine andere Lösung der Aufgabe besteht darin, eine Anordnung mit einer Prozessoreinheit anzugeben, welche Prozessoreinheit derart eingerichtet ist, daß folgende Schritte durchführbar sind:

a) aus der gesprochenen Sprache wird ein digitalisiertes Sprachsignal bestimmt;,

b) auf dem digitalisierten Sprachsignal erfolgt eine Signalanalyse, woraus Merkmalsvektoren zur Beschreibung des digitalisierten Sprachsignals hervorgehen;

c) eine globale Suche zur Abbildung der Merkmalsvektoren erfolgt auf eine in modellierter Form vorliegende Sprache, wobei Phoneme der Sprache durch ein modifiziertes Hidden-Markov-Modell und jeder Zustand des Hidden-Markov-Modells durch eine Wahrscheinlichkeitsdichtefunktion beschreibbar ist;

d) die wird Wahrscheinlichkeitsdichtefunktion durch Veränderung des Wortschatzes angepaßt, indem die Wahrscheinlichkeitsdichtefunktion in eine erste Wahrscheinlichkeitsdichtefunktion und in eine zweite Wahrscheinlichkeitsdichtefunktion aufgespalten wird;

e) von der globalen Suche wird eine erkannte Wortfolge bereitgestellt.

Diese Anordnung ist insbesondere geeignet zur Durchführung des erfindungsgemäßen Verfahrens oder einer seiner vorstehend erläuterten Weiterbildungen.

Ausführungsbeispiele der Erfindung werden nachfolgend anhand der Zeichnung dargestellt und erläutert.

Es zeigt

Fig.1 eine Anordnung bzw. ein Verfahren zur Erkennung gesprochener Sprache.

In Figur 1 sind eine Anordnung bzw. ein Verfahren zur Erkennung gesprochener Sprache dargestellt. Zur Erläuterung der nachstehend verwendeten Begriffe sei auf die Beschreibungseinleitung verwiesen.

Ein digitalisiertes Sprachsignal 101 wird in einer Signalanalyse 102 einer Fouriertransformation 103 mit

anschließender Merkmalsextraktion 104 unterzogen. Die Merkmalsvektoren 105 werden an ein System zur globalen Suche 106 übermittelt. Die globale Suche 106 berücksichtigt sowohl ein akustisches Modell 107 als auch ein linguistisches Modell 108 zur Bestimmung der erkannten Wortfolge 109. Aus dem digitalisierten Sprachsignal 101 geht somit die erkannte Wortfolge 109 hervor.

In dem akustischen Modell 107 wird das Phoneminventar anhand von Hidden-Markov-Modellen nachgebildet.

Eine Wahrscheinlichkeitsdichtefunktion eines Zustands des Hidden-Markov-Modells wird durch eine Aufsummierung einzelner Gaußscher Moden angenähert. Eine Mode ist insbesondere eine Gaußglocke. Durch Aufsummierung mehrerer Moden entsteht eine Mischung einzelner Gaußglocken und damit eine Modellierung der Emissionswahrscheinlichkeitsdichtefunktion. Anhand eines statistischen Kriteriums wird entschieden, ob der zu erkennende Wortschatz des Spracherkenners durch das Hinzufügen weiterer Moden verbessert modelliert werden kann. Im Fall der vorliegenden Erfindung wird dies insbesondere bei Erfüllung des statistischen Kriteriums durch inkrementelles Aufspalten bereits existierender Moden erreicht.

Die Entropie ist definiert durch

$$H_p = -\int_{-\infty}^{\infty} p(\vec{x}) \log_2 p(\vec{x})\, d\vec{x} \qquad (1)$$

unter der Annahme, daß $p(\vec{x})$ eine Gauß-Verteilung mit einer diagonalen Kovarianzmatrix ist, also

$$p(\vec{x}) = \mathcal{N}(\vec{\mu}, \sigma_n) = \frac{1}{\sqrt{(2\pi)^N}} \frac{1}{\prod_n \sigma_n} \cdot \exp\left(-\frac{1}{2} \sum_n \frac{(x_n - \mu_n)^2}{\sigma_n^2} \right) \qquad (2)$$

erhält man

$$H_p = \sum_{n=1}^{N} \log_2 \sqrt{2\pi e}\ \sigma_n \qquad (3),$$

wobei

μ den Erwartungswert,
σ_n die Streuung für jede Komponente n und
N die Dimension des Merkmalsraums

bezeichnen.

Die wahre Verteilung $p(\vec{x})$ ist nicht bekannt. Sie wird insbesondere als Gaußverteilung angenommen. Im akustischen Modell wird die Wahrscheinlichkeit $p(\vec{x})$ anhand von Stichproben angenähert mit

$$\hat{p}(\vec{x}) = \mathcal{N}\left(\vec{\hat{\mu}}, \sigma_n\right),$$

wobei

$$\vec{\hat{\mu}} = \frac{1}{L} \sum_{l=1}^{L} \vec{x}_l$$

einen Mittelwert über L Beobachtungen darstellt. Die korrespondierende Entropie als Funktion von $\hat{\mu}$ ist gegeben durch

$$H_{\hat{p}}(\hat{\mu}) = -\int_{-\infty}^{\infty} p(\vec{x}) \log_2 \hat{p}(\vec{x})\ d\vec{x} \qquad (4),$$

was schließlich zu

$$H_{\hat{p}}(\hat{\mu}) = H_p + \sum_{n=1}^{N} \frac{(\mu_n - \hat{\mu}_n)^2}{\sigma_n^2} \log_2 \sqrt{e} \qquad (5)$$

führt.

Der Erwartungswert $E\left\{(\mu_n - \hat{\mu}_n)^2\right\}$ beträgt $\frac{1}{L}\sigma_n^2$, so daß der Erwartungswert von $H_{\hat{p}}(\hat{\mu})$ gegeben ist als

$$H_{\hat{p}} = E\left\{H_{\hat{p}}(\hat{\mu})\right\} = H_p + \frac{N}{L}\log_2\sqrt{e} \qquad (6).$$

Für die Entropie einer Mode, die mit einer Gauß-Verteilung mit einer diagonalen Kovarianzmatrix bestimmt wird, ergibt sich also Gleichung (3). Der Prozeß wird nun mit einer Schätzung angenähert. Die Entropie des angenäherten Prozesses ergibt sich zu

$$\hat{H} = H + \frac{N}{L}\log_2\sqrt{e} \qquad (7).$$

Je größer die Anzahl L der Stichproben ist, um so besser wird die Abschätzung und um so mehr nähert sich die geschätzte Entropie $\hat{H}$ der wahren Entropie H an.

Es soll nun

$$p(\vec{x}) = \mathcal{N}\left(\vec{\tilde{\mu}}, \sigma_n\right) \qquad (8)$$

die aufzuteilende Mode sein. Ferner wird angenommen, daß die zwei Gauß-Verteilungen, die durch den Aufteilungsprozeß entstehen, identische Standardabweichungen σ^s haben und gleich gewichtet sind. Dies ergibt

$$\hat{p}^s(\vec{x}) = \frac{1}{2}\mathcal{N}\left(\vec{\tilde{\mu}}_1^s, \sigma^s\right) + \frac{1}{2}\mathcal{N}\left(\vec{\tilde{\mu}}_2^s, \sigma^s\right) \qquad (9).$$

Unter der Annahme, daß $\mu_1 \approx \hat{\mu}_1$, $\mu_2 \approx \hat{\mu}_2$ und daß μ_1 ausreichend weit entfernt von μ_2 ist, ergibt sich die Entropie der aufgespaltenen Wahrscheinlichkeitsdichtefunktion jeweils zu

$$\hat{H}^S = 1 - \sum_{n=1}^{N} \log_2 \sqrt{2\pi e}\, \sigma_n^S + \frac{1}{2}\left(\log_2 \sqrt{e}\, \frac{N}{L_1} + \log_2 \sqrt{e}\, \frac{N}{L_2}\right) \quad (10).$$

Als Aufteilungskriterium wird eine Verminderung der Entropie durch den Aufspaltungsvorgang gefordert, also

$$\hat{H} - \hat{H}^S > C \quad (11),$$

wobei C (mit $C > 0$) eine Konstante ist, die den gewünschten Abfall der Entropie darstellt. Wird

$$\frac{L}{2} = L_1 = L_2 \quad (12)$$

angenommen, so ergibt sich hierdurch

$$\sum_{n=1}^{N} \log_2 \frac{\sigma_n}{\sigma_n^S} > \log_2 \sqrt{e}\, \frac{N}{L} + 1 + C \quad (13).$$

Eine Möglichkeit, die Lage der Mittelpunkte der beiden neuen Moden zu bestimmen, wird im folgenden aufgezeigt. Eine bevorzugte Vorgabe besteht darin, daß Kriterium zum Aufspalten zu erfüllen. In dem angeführten Beispiel wird $\hat{\mu}_1^S$ der Wert von $\hat{\mu}$ zugewiesen. $\hat{\mu}_2^S$ erhält eine Maximum-Likelihood-Schätzung derjenigen Beobachtungen, die im Viterbi-Pfad auf $\hat{\mu}$ abgebildet werden. Diese Bestimmungen zeigen lediglich eine Möglichkeit auf, ohne daß eine Einschränkung des vorgestellten Verfahrens auf diese Möglichkeit beabsichtigt ist.

Die folgenden Schritte der Beispielanwendung zeigen die Einbettung in eine Anordnung zur Spracherkennung bzw. ein Verfahren zur Spracherkennung.

Schritt 1: Initialisierung: $\vec{\mu}_1^S = \vec{\mu}$, $\vec{\mu}_2^S = \vec{\mu}$.

Schritt 2: Erkennen der Äußerung , Analysieren des Viterbi-Pfads;

Schritt 3: Für jeden Zustand und für jede Mode des Viterbi-Pfades:

Schritt 3.1: Bestimme σ_n;

Schritt 3.2: Bestimme L_2 auf Grundlage derjenigen Beobachtungen, die näher an $\vec{\mu}_2^s$ als an $\vec{\mu}_1^s$ liegen und setze $L = L_2$. Falls $\vec{\mu}_2^s$ und $\vec{\mu}_1^s$ identisch sind, so ordne die zweite Hälfte der Merkmalsvektoren $\vec{\mu}_2^s$ und die erste Hälfte der Merkmalsvektoren $\vec{\mu}_1^s$ zu.

Schritt 3.3: Bestimme σ_n^s entsprechend auf Grundlage der L_2-Äußerungen;

Schritt 3.4: Ermittle $\vec{\mu}_2^s$ neu auf Grundlage des Mittelwerts derjenigen Beobachtungen, die näher an $\vec{\mu}_2^s$ als an $\vec{\mu}_1^s$ liegen;

Schritt 3.5: Werte Aufteilungskriterium nach Gleichung (13) aus;

Schritt 3.6: Falls Aufteilungskriterium nach Gleichung (13) positiv ist, generiere zwei neue Moden mit den Mittelpunkten $\vec{\mu}_1^s$ und $\vec{\mu}_2^s$.

Schritt 4: Gehe zu Schritt 2.

Im Rahmen dieses Dokuments wurden folgende Veröffentlichungen zitiert:

[1] N. Haberland et al.: "Sprachunterricht - Wie funktioniert die computerbasierte Spracherkennung?", c't - Magazin für Computertechnik - 5/1998, Heinz Heise Verlag, Hannover, 1998, Seiten 120 bis 125.

[2] C. H. Lee et al.: "Speaker Adaptation Based on MAP Estimation of HMM Parameters"; Proc. IEEE Intern. Conference on Acoustics, Speech and Signal Processing, Seiten II-588 bis II-561.

Patentansprüche

1. Verfahren zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache durch einen Rechner,
   a) bei dem aus der gesprochenen Sprache ein digitalisiertes Sprachsignal bestimmt wird,
   b) bei dem auf dem digitalisierten Sprachsignal eine Signalanalyse durchgeführt wird, woraus Merkmalsvektoren zur Beschreibung des digitalisierten Sprachsignals hervorgehen,
   c) bei dem eine globale Suche zur Abbildung der Merkmalsvektoren auf eine in modellierter Form vorliegende Sprache durchgeführt wird, wobei Phoneme der Sprache durch ein modifiziertes Hidden-Markov-Modell und jeder Zustand des Hidden-Markov-Modells durch eine Wahrscheinlichkeitsdichtefunktion beschrieben wird,
   d) bei dem die Wahrscheinlichkeitsdichtefunktion durch Veränderung des Wortschatzes angepaßt wird, indem die Wahrscheinlichkeitsdichtefunktion in eine erste Wahrscheinlichkeitsdichtefunktion und in eine zweite Wahrscheinlichkeitsdichtefunktion aufgespalten wird,
   e) bei dem von der globalen Suche eine erkannte Wortfolge bereitgestellt wird.

2. Verfahren nach Anspruch 1,
   bei dem die Wahrscheinlichkeitsdichtefunktion in die erste und in die zweite Wahrscheinlichkeitsdichtefunktion aufgespalten wird, falls der Abfall eines Entropiewertes unterhalb einer vorgegebenen Schranke liegt.

3. Verfahren nach Anspruch 1 oder 2,
   bei dem die Veränderung des Wortschatzes zur Laufzeit des Verfahrens durchgeführt wird.

4. Verfahren nach einem der vorhergehenden Ansprüche,
   bei dem die Veränderung des Wortschatzes bedingt ist

durch Hinzufügen eines Wortes zum Wortschatz oder bei dem sich Aussprachegewohnheiten eines Sprechers ändern.

5. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die erste Wahrscheinlichkeitsdichtefunktion und die zweite Wahrscheinlichkeitsdichtefunktion jeweils mindestens eine Gaußverteilung umfassen.

6. Verfahren nach Anspruch 5, bei dem für die erste Wahrscheinlichkeitsdichtefunktion und für die zweite Wahrscheinlichkeitsdichtefunktion gleiche Standardabweichungen, ein erster Mittelwert der ersten Wahrscheinlichkeitsdichtefunktion und ein zweiter Mittelwert der zweiten Wahrscheinlichkeitsdichtefunktion ermittelt werden, wobei der erste Mittelwert von dem zweiten Mittelwert verschieden ist.

7. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Aufspaltung mehrfach durchgeführt wird.

8. Anordnung zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache mit einer Prozessoreinheit, die derart eingerichtet ist, daß

a) aus der gesprochenen Sprache ein digitalisiertes Sprachsignal bestimmbar ist,

b) auf dem digitalisierten Sprachsignal eine Signalanalyse durchführbar ist, woraus Merkmalsvektoren zur Beschreibung des digitalisierten Sprachsignals hervorgehen,

c) eine globale Suche zur Abbildung der Merkmalsvektoren auf eine in modellierter Form vorliegende Sprache erfolgt, wobei Phoneme der Sprache durch ein modifiziertes Hidden-Markov-Modell und jeder Zustand des Hidden-Markov-Modells durch eine Wahrscheinlichkeitsdichtefunktion beschreibbar ist,

d) die Wahrscheinlichkeitsdichtefunktion durch Veränderung des Wortschatzes angepaßt wird, indem die Wahrscheinlichkeitsdichtefunktion in eine erste Wahrscheinlichkeitsdichtefunktion und in eine zweite Wahrscheinlichkeitsdichtefunktion aufgespalten wird,

e) von der globalen Suche eine erkannte Wortfolge bereitgestellt wird.

Zusammenfassung

Anordnung und Verfahren zur Erkennung eines vorgegebenen Wortschatzes in gesprochener Sprache durch einen Rechner

Bei der Spracherkennung werden Phoneme einer Sprache durch ein Hidden-Markov-Modell modelliert, wobei jeder Zustand des Hidden-Markov-Modells durch eine Wahrscheinlichkeitsdichtefunktion beschrieben wird. Zur Spracherkennung eines veränderten Wortschatzes wird die Wahrscheinlichkeitsdichtefunktion in eine erste und eine zweite Wahrscheinlichkeitsdichtefunktion aufgespalten. Dadurch wird es möglich, Veränderungen der Sprachgewohnheit eines Sprechers zu kompensieren oder ein neues Wort dem Wortschatz des Spracherkenners hinzuzufügen und dabei sicherzustellen, daß dieses neue Wort mit ausreichender Güte von den bereits im Spracherkenner vorhandenen Wörtern unterschieden und somit erkannt wird.

Figur 1

Fig 1 GR 98 P 1687